

weld rigid grasp soft sort transport highway intersection inspect pallet barcode scan carry tool shiny kitchen living room robot operate book h  
ognition cut greenhouse clothing item street intersection vehicle doorway appliance wheeled-cart pliable pallet factory floor barcode pour clea  
chair pour rough tablet office subway station locate door handle cup open/close screw/unscrew paint pick/place follow charge climb clear  
navigate barcode reflective scan texture transport highway intersection inspect pallet barcode scan carry tool shiny kitchen living room robot  
slippery brick wall shiny climb push pull navigate transport scan chair table couch avoid obstacle high friction grainy texture construction  
absorbent lab environment tool grip pack school classroom solid state pliable spongy office subway station locate door handle rough surface grip edge  
landmark fluid construction site factory floor subway station pour recognize school classroom navigate align pick/place follow screw wheeled cart cup s  
street intersection pour climb navigate grasp sort recognize cut garden hotel lobby transport scan clean push/pull stir paint screw/unscrew weld inspect  
open/close pick/place follow carry clean sort stir navigate sticky suction shiny grip edge transport barcode shiny climb push pull navigate transport scar  
table couch avoid obstacle high friction grainy texture retain friction slippery check barcode heat-sensitive thermal properties compression under load c  
touch object's affordances provide leverage bending resistance act on signal illuminate switch toggle engage safety feature reinforced corner delicate s  
load-bearing  
feedback su  
contour ma  
side handle  
stability erg  
malleable fo  
port haptic  
rolling resis  
depth mark  
mounting s  
noise modu  
adhesive ba  
limit beveled  
edge retractive  
inax object symme  
ancnoring slot ergonomic grip tnermal expansion cut-resistant modular joint waterproof seal modula  
interface flexural strength weighting agent radiation shielding latching system visual marker texture contrast arm extension torque specification spill-pr  
containment crumple zone load distribution fracture toughness quick release catch maximal load capacity display surface bending flexibility optical sen  
interface payload alignment incremental adjustment sliding slot roughened grip strain relief tether point load dispersion lubrication reservoir sealed com  
natural finish scratch resistance permanent magnet castor wheel friction lining inertia brake soft grip touch sensor panel displaceable volume broadcast  
decompression limit vacuum seal electrical grounding lever action suspension hook alignment pin illuminated display safety interlock adjustable clamp  
dampening balance actuator detachable joint air flow regulation splinter guard cantilever support biodegradable material wear-resistant surface reversib  
function tensile load buffer zone hatch opening clamping pressure signal conduit polarized filter display brightness limiter aerodynamic profile sealing g  
energy damping recoil mechanism return spring counterweight ballast dispersible medium abrasion-resistant coating sensitivity adjustment temperatur  
compensation grip contour electrical continuity axle support partitioned space luminescent signal handle traction floating hinge breathable fabric electric  
insulator swing limit motion path guide strain gauge sensor calibrated scale non-slip footing articulated linkage deployable support gradient measure im  
cushion variable aperture coupling adapter self-lubricating surface swivel base magnetic latch folding mechanism quick disconnect system light-diffusin  
ighted base fluted grip sound insulation photoelectric cell collision buffer angle brace damping coefficient motion gu  
isolation pneumatic support sliding resistance tensioning device encapsulated system multi-axis rotation socket  
ed platform notch filter rigidity modulator clasp locking tactile indicator heat dispersion fin illumination pattern an  
quid retention piston drive snap-fit connection non-marking base reversible linkage telescoping arm adjustable s



# DEEP Rob

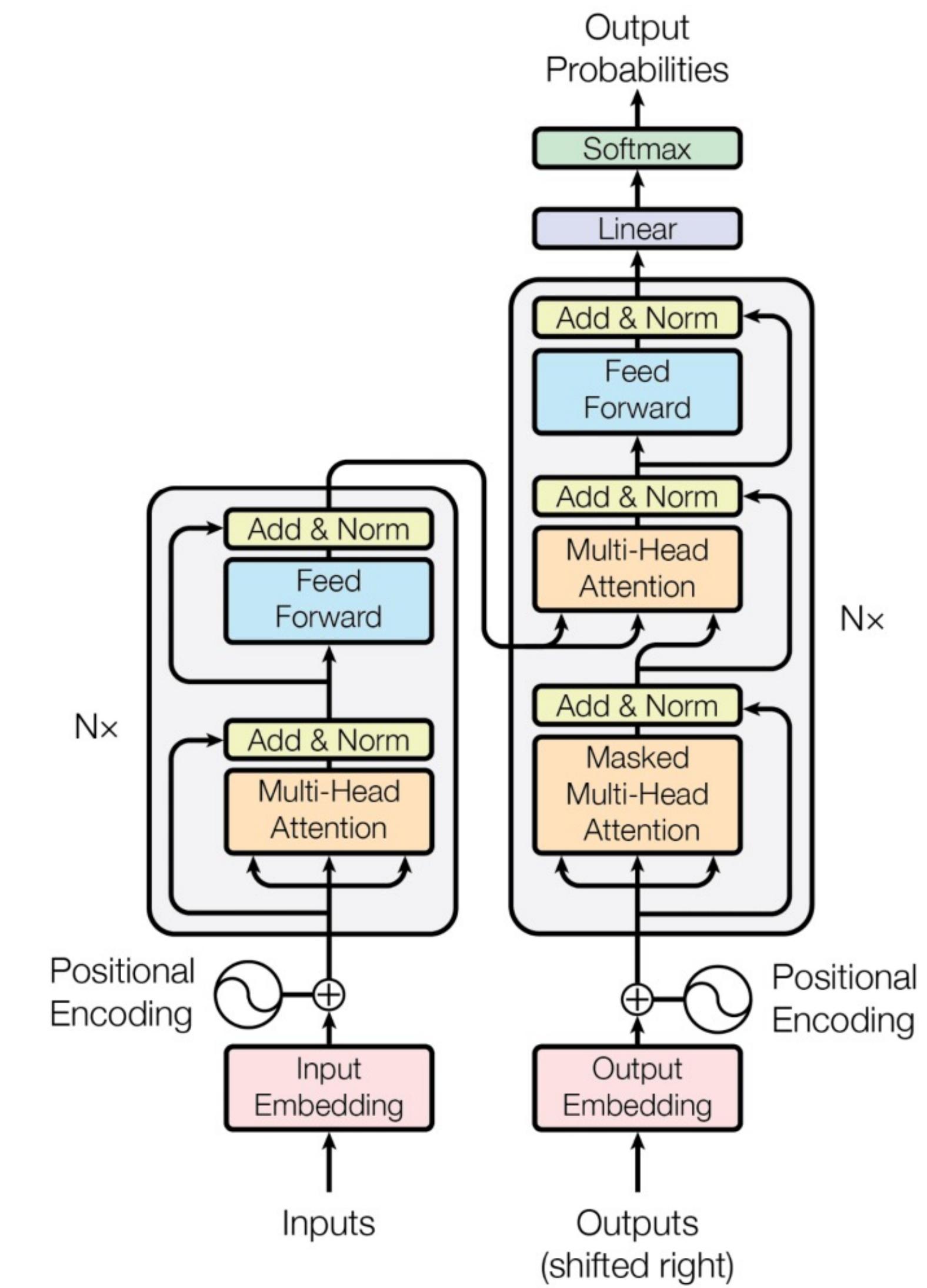
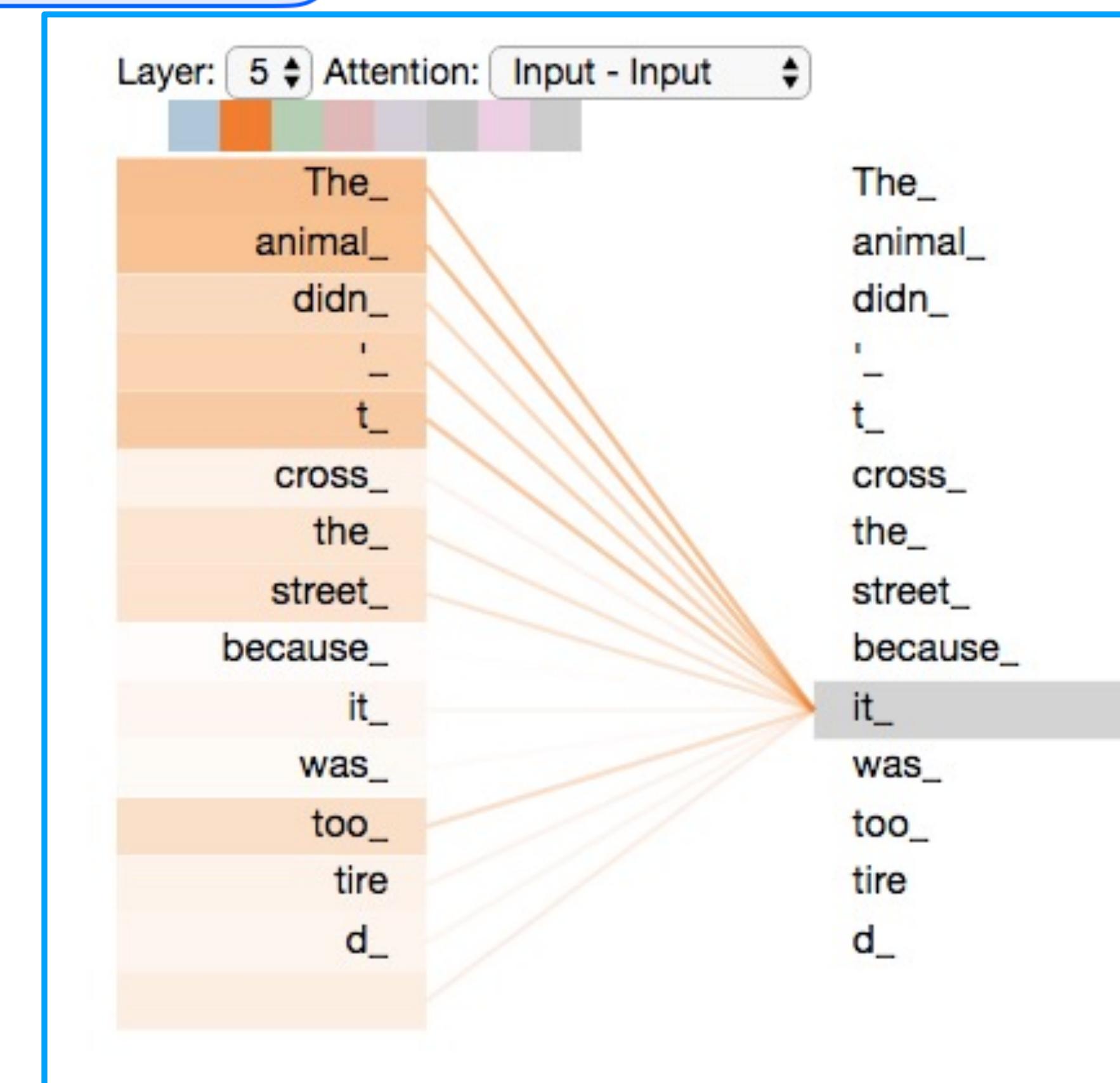
## Lecture 16 Language Models University of Michigan I Department of Robotics



DEEP Rob



# Transformers (review)





# Transformers Mid-2017

**Input** – input tokens

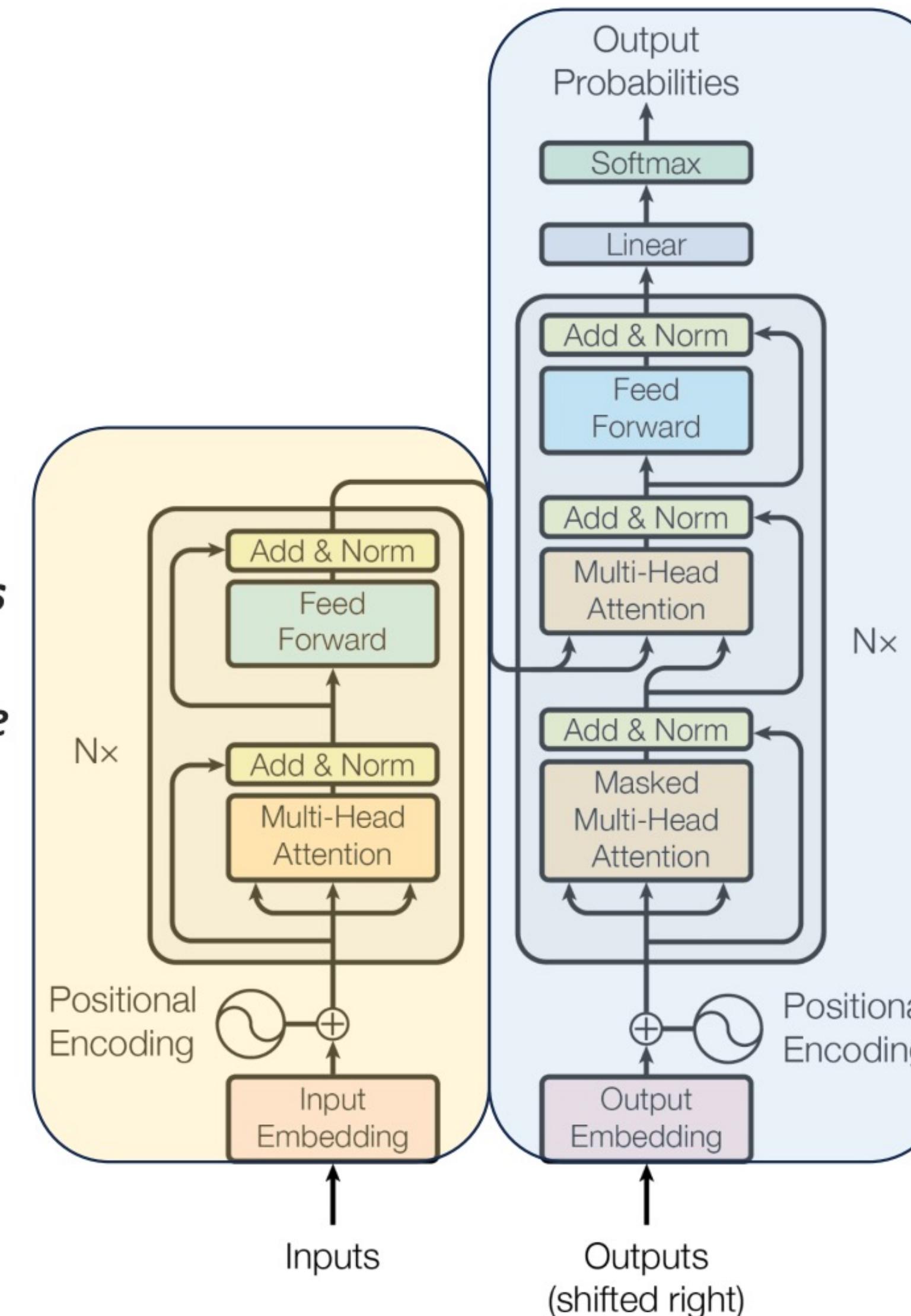
**Output** – hidden states

**Model can see all timesteps**

**Does not usually output tokens, so no inherent auto-regressivity**

***Can also be adapted to generate tokens by appending a module that maps hidden state dimensionality to vocab size***

## Representation



**Input** – output tokens and hidden states\*

**Output** – output tokens

**Model can only see previous timesteps**

**Model is auto-regressive with previous timesteps' outputs**

***Can also be adapted to generate hidden states by looking before token outputs***

## Generation

<https://deeplearning.cs.cmu.edu/F23/document/slides/lec19.transformersLLMs.pdf>

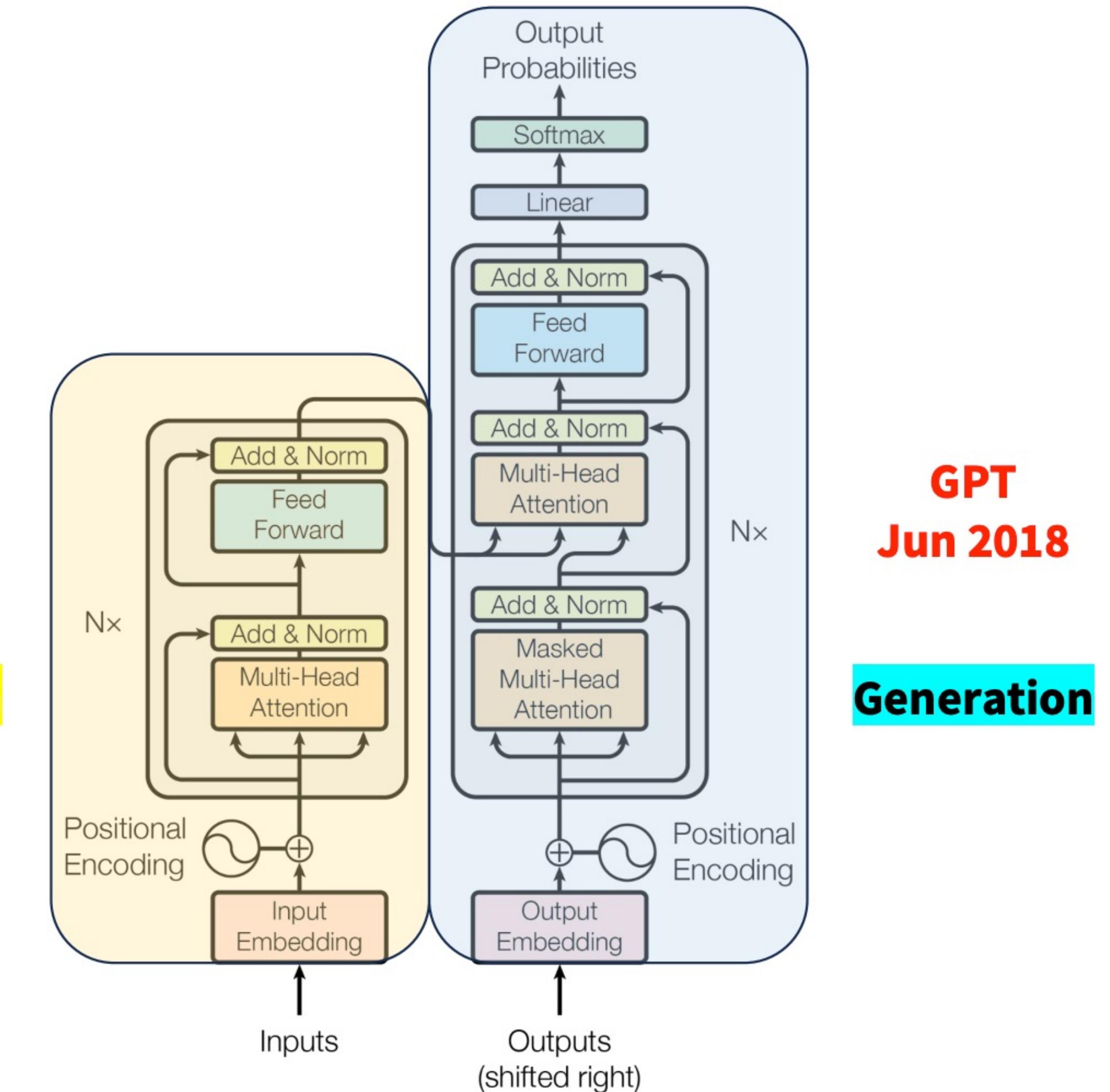




2018-  
LLM Era

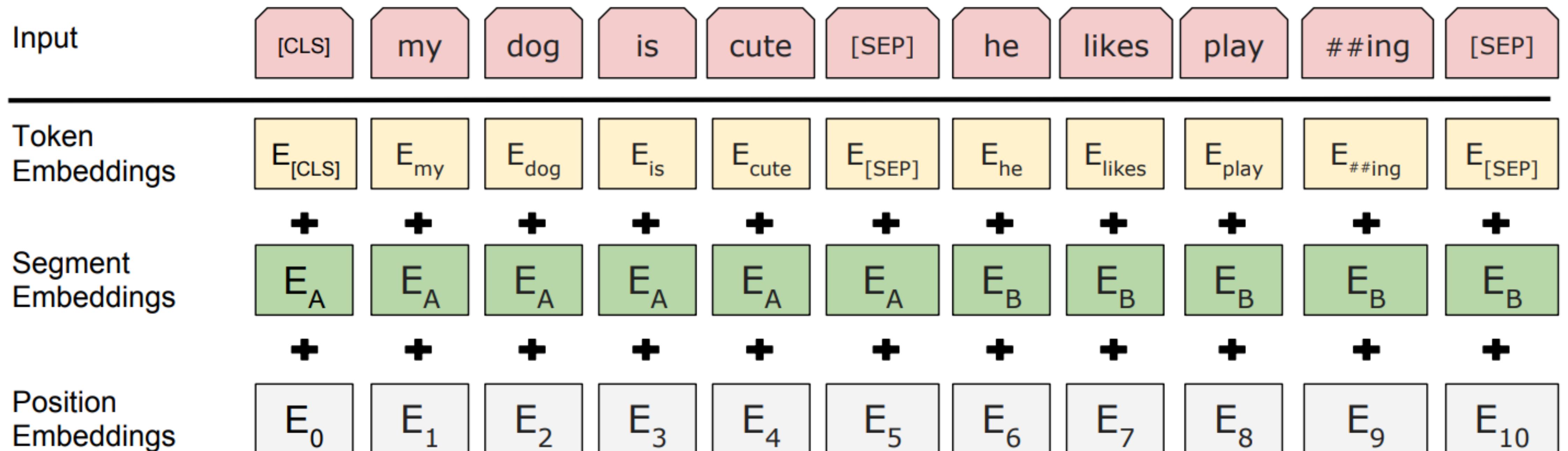
**BERT**  
**Oct 2018**

**Representation**



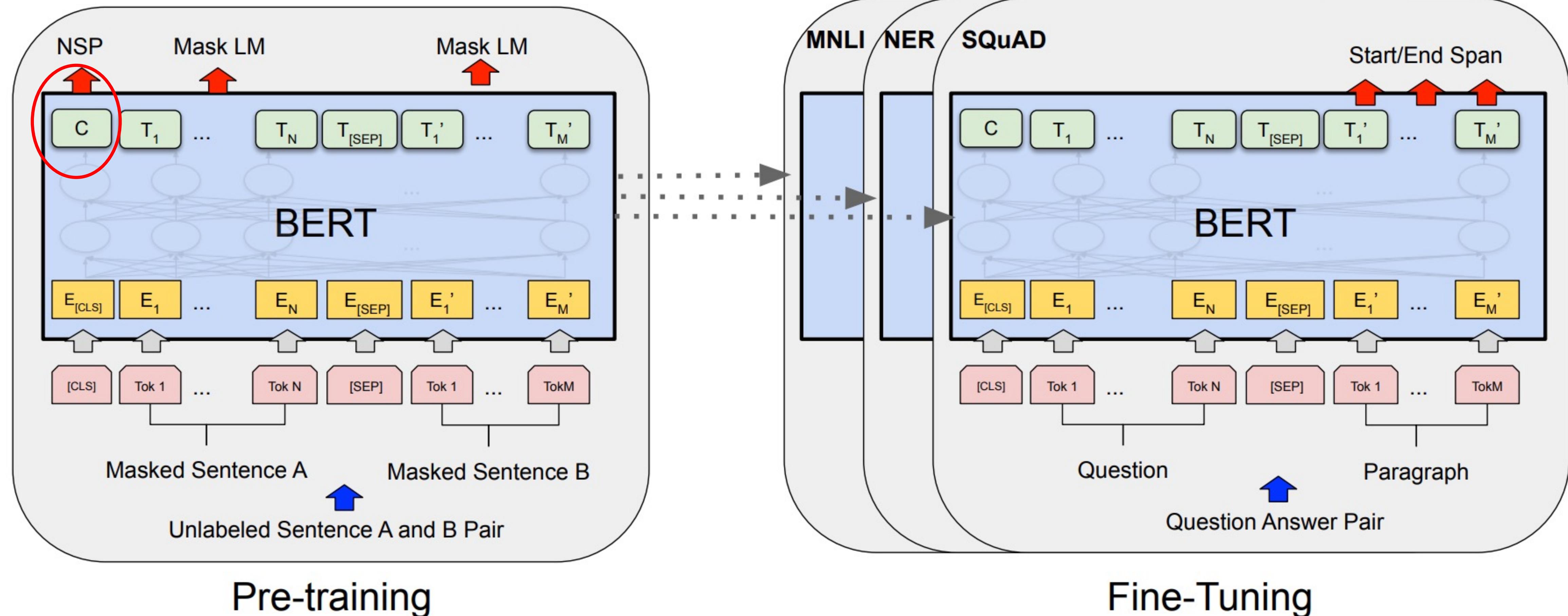


# BERT Input representations





# BERT: Bidirectional Encoder Representations from Transformers



Pre-training

Fine-Tuning



# BLEU metric $\in [0,1]$

---

- Bilingual Evaluation Understudy
- <https://huggingface.co/spaces/evaluate-metric/bleu>
- R (reference): human expert
- C (candidate): produced by translation system (e.g., a Transformer)



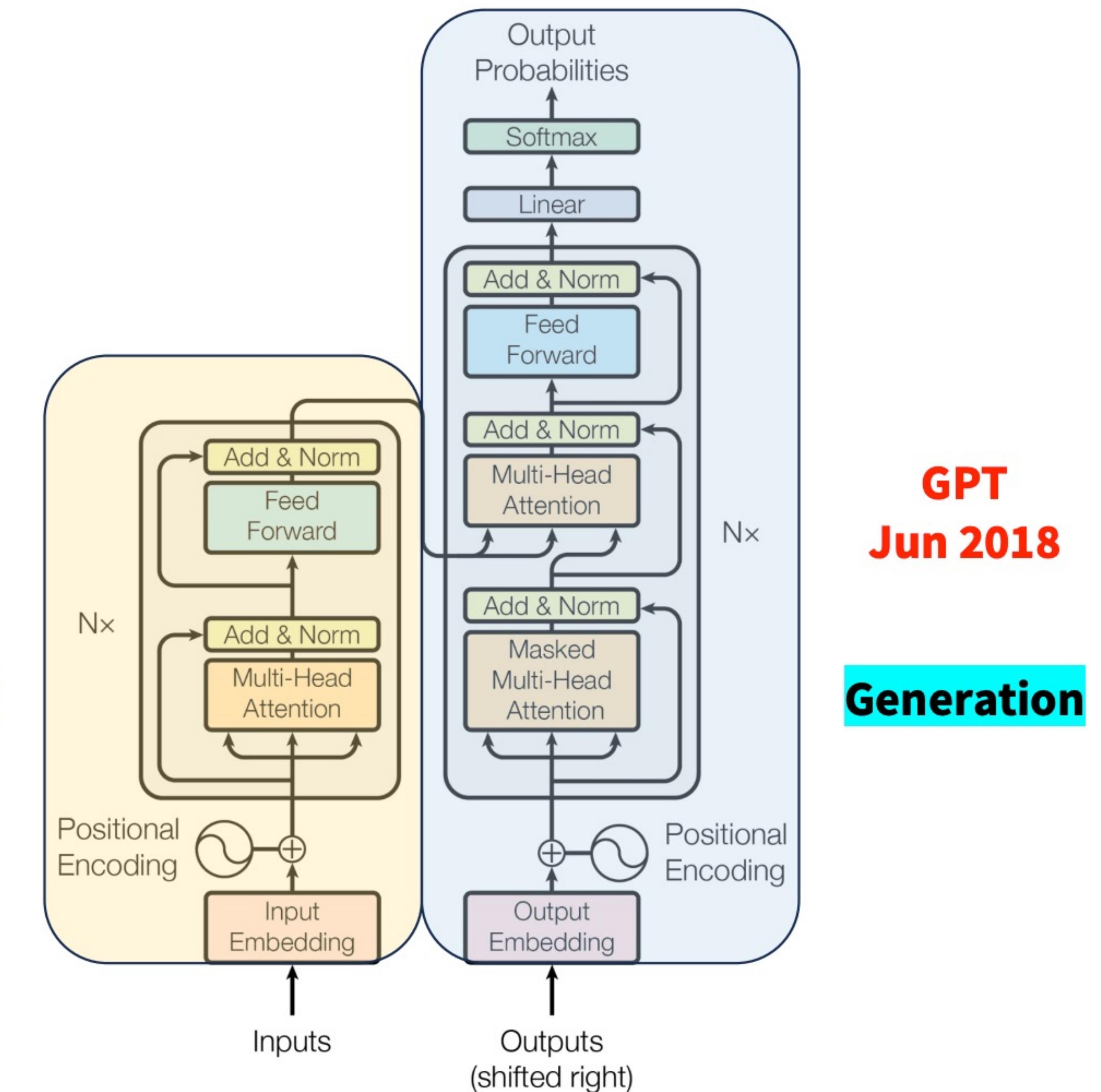
## 2018- LLM Era

### Corpus

- English Wikipedia  
- 2,500 million words
- Book Corpus - 800 million words

**BERT**  
**Oct 2018**

**Representation**



**GPT**  
**Jun 2018**

**Generation**



# CLIP

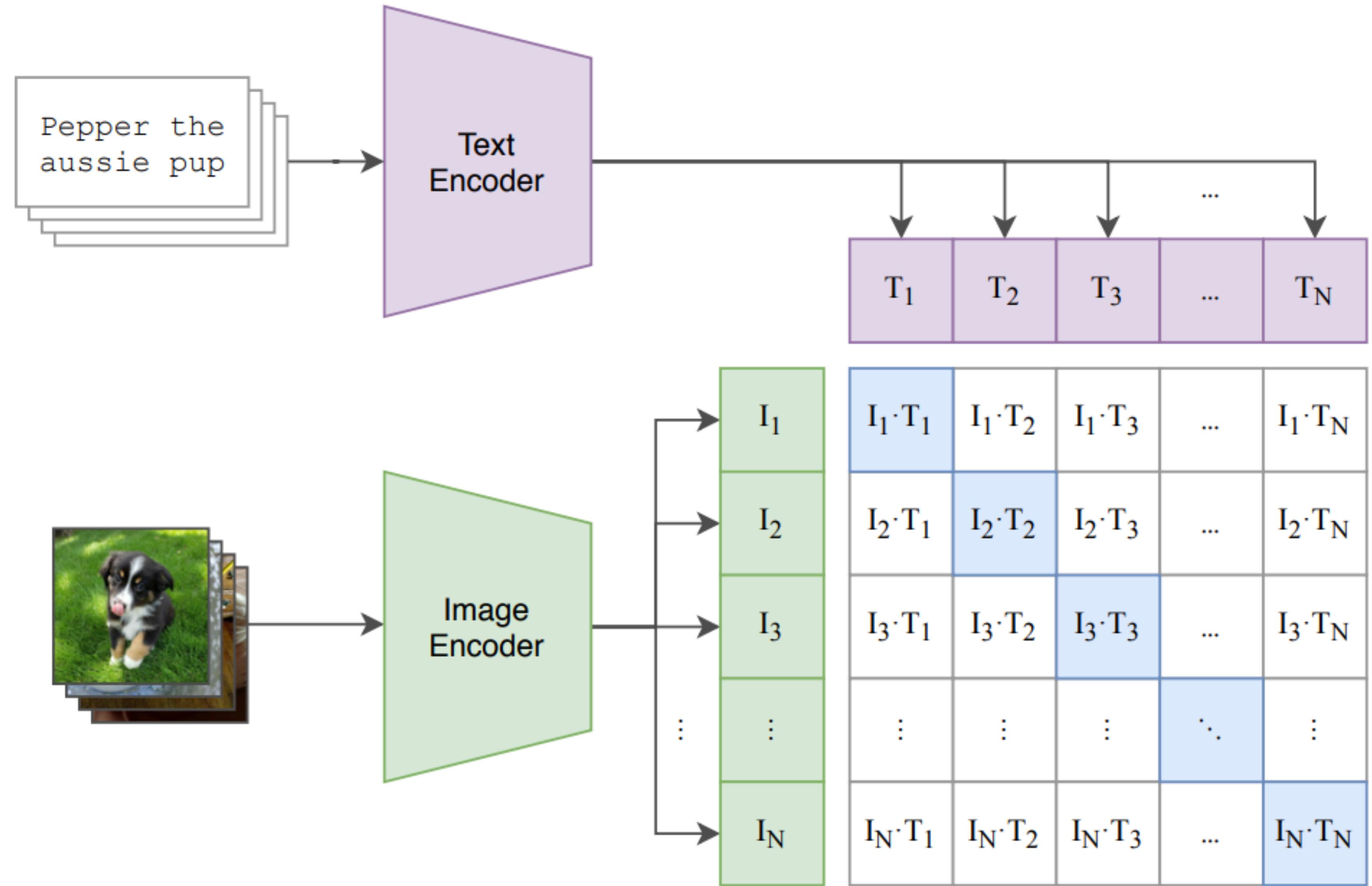
---

- CLIP (*Contrastive Language–Image Pre-training*)
- learning **visual** representations from **natural language** supervision



# CLIP

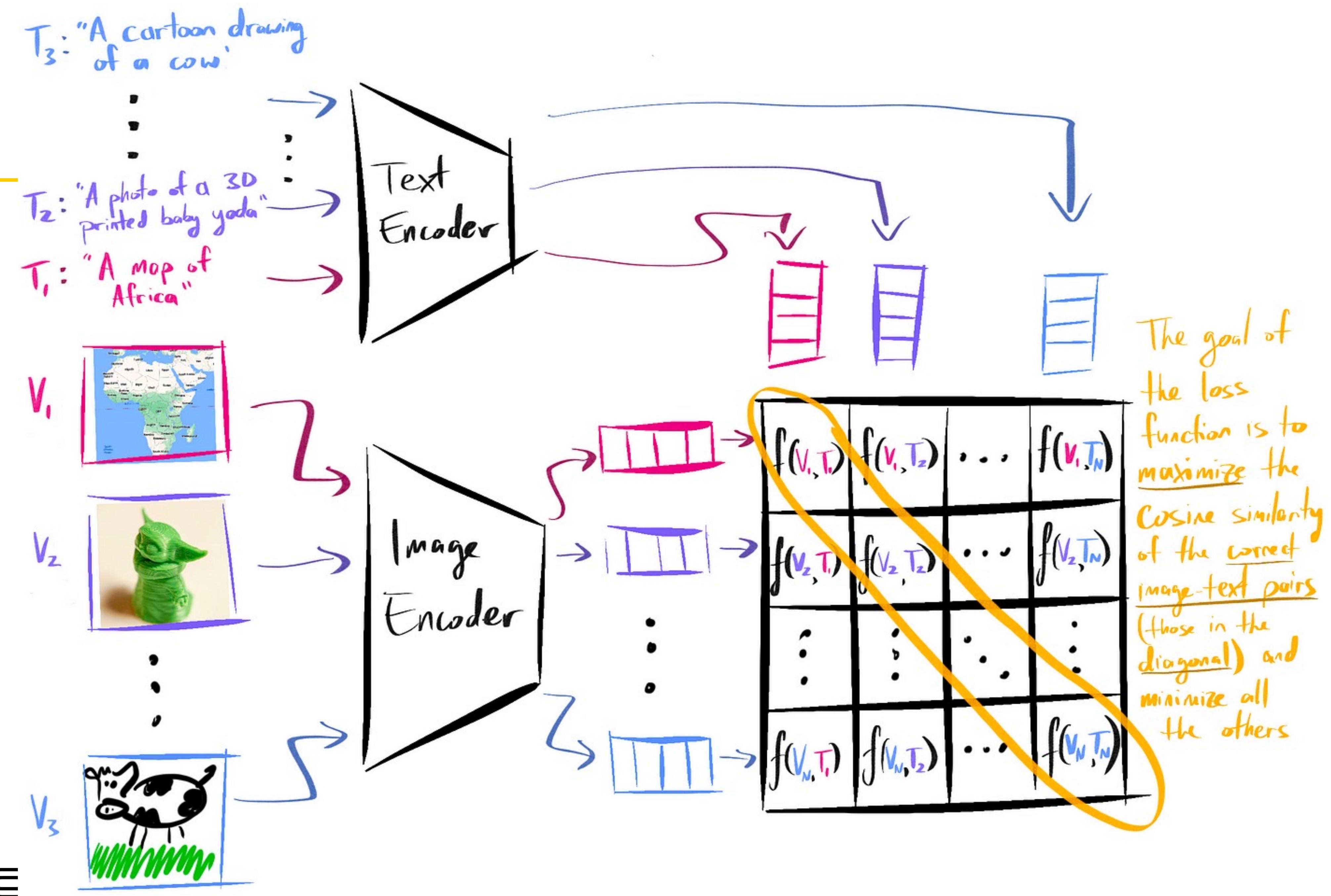
## (1) Contrastive pre-training





# CLIP

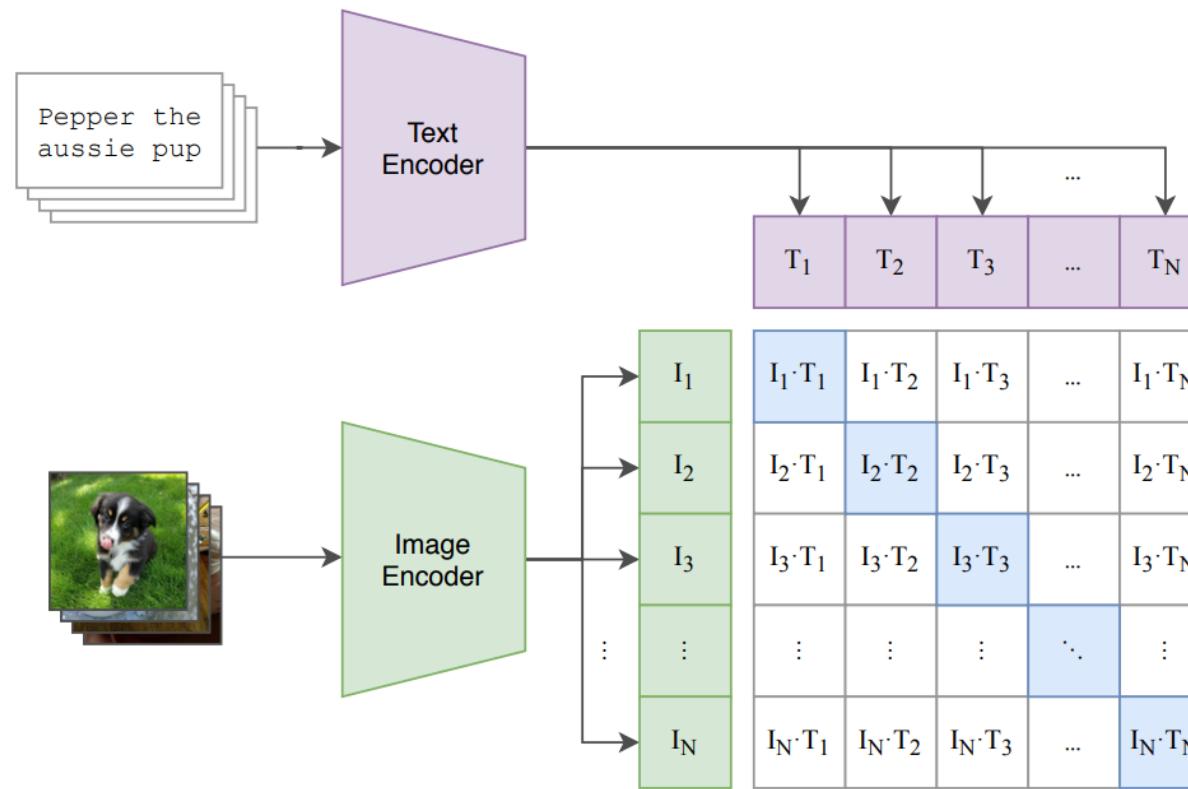
(example)



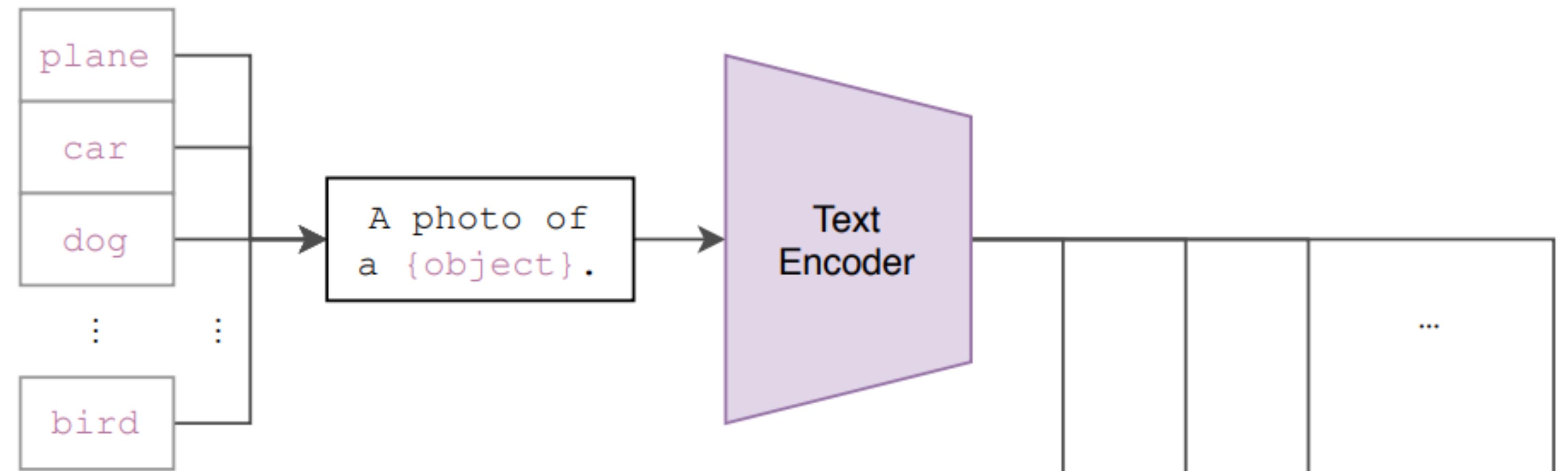


# CLIP

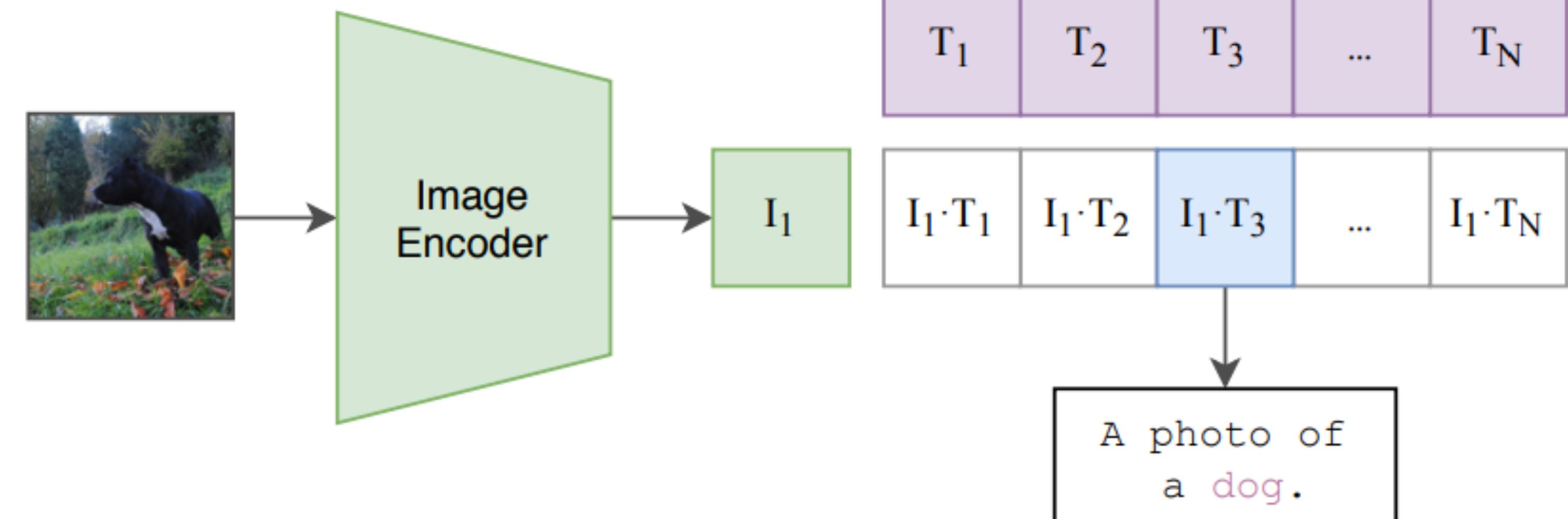
## (1) Contrastive pre-training



## (2) Create dataset classifier from label text



## (3) Use for zero-shot prediction





# CLIP

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, 1]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

*Figure 3.* Numpy-like pseudocode for the core of an implementation of CLIP.



# CLIP

---

- learning **visual** representations from **natural language** supervision
- Pre-trained model, NOT a generative model
- Advantage:
  - does not need task-specific training data
  - bridging two modalities
- Limitations:
  - abstract or systematic tasks, complex tasks (e.g., predicting “nearest”, counting)
  - poor generalization on images not covered in pre-training



# LERF

- LERF:  
Language  
Embedded  
Radiance  
Field
- *view-  
independent*

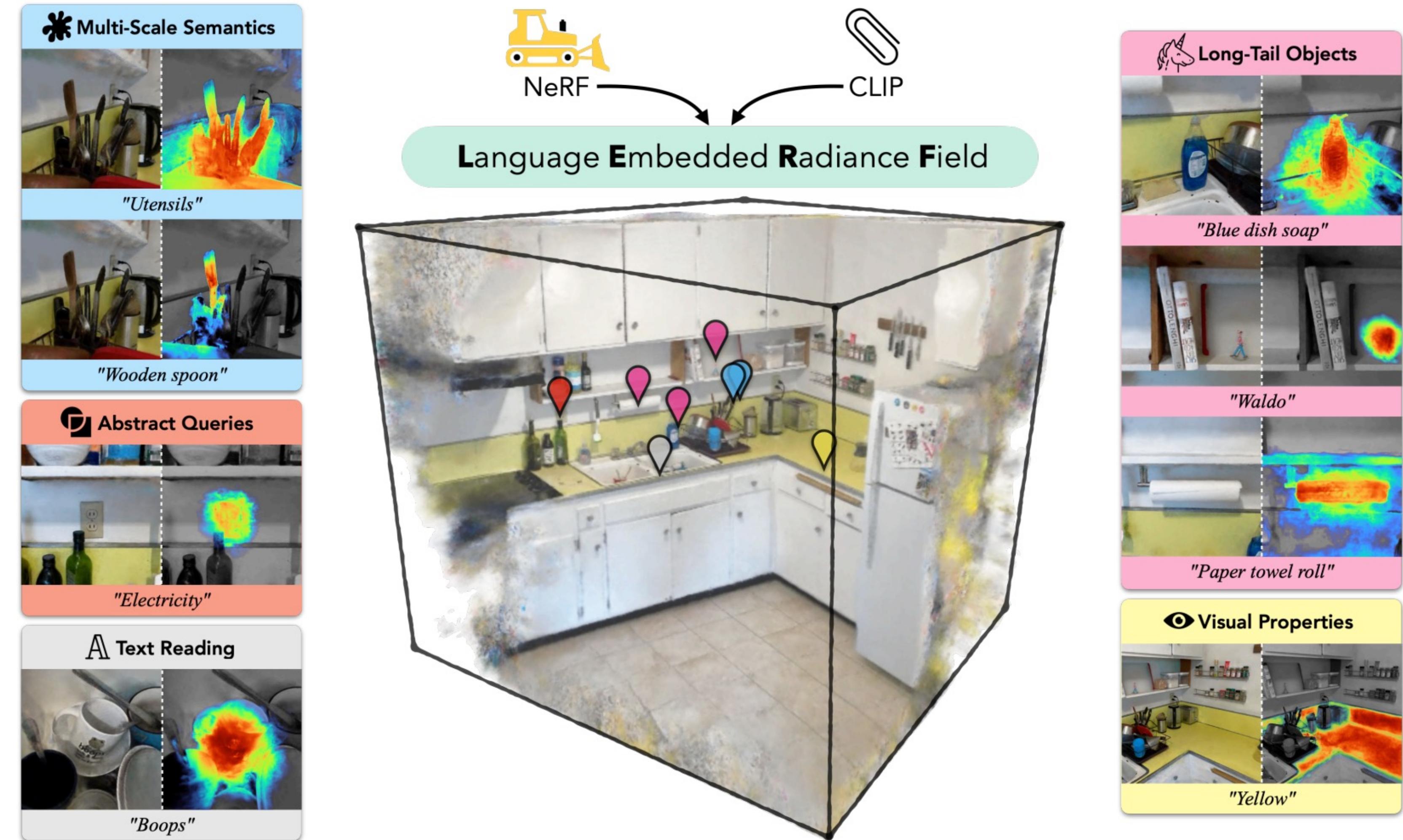
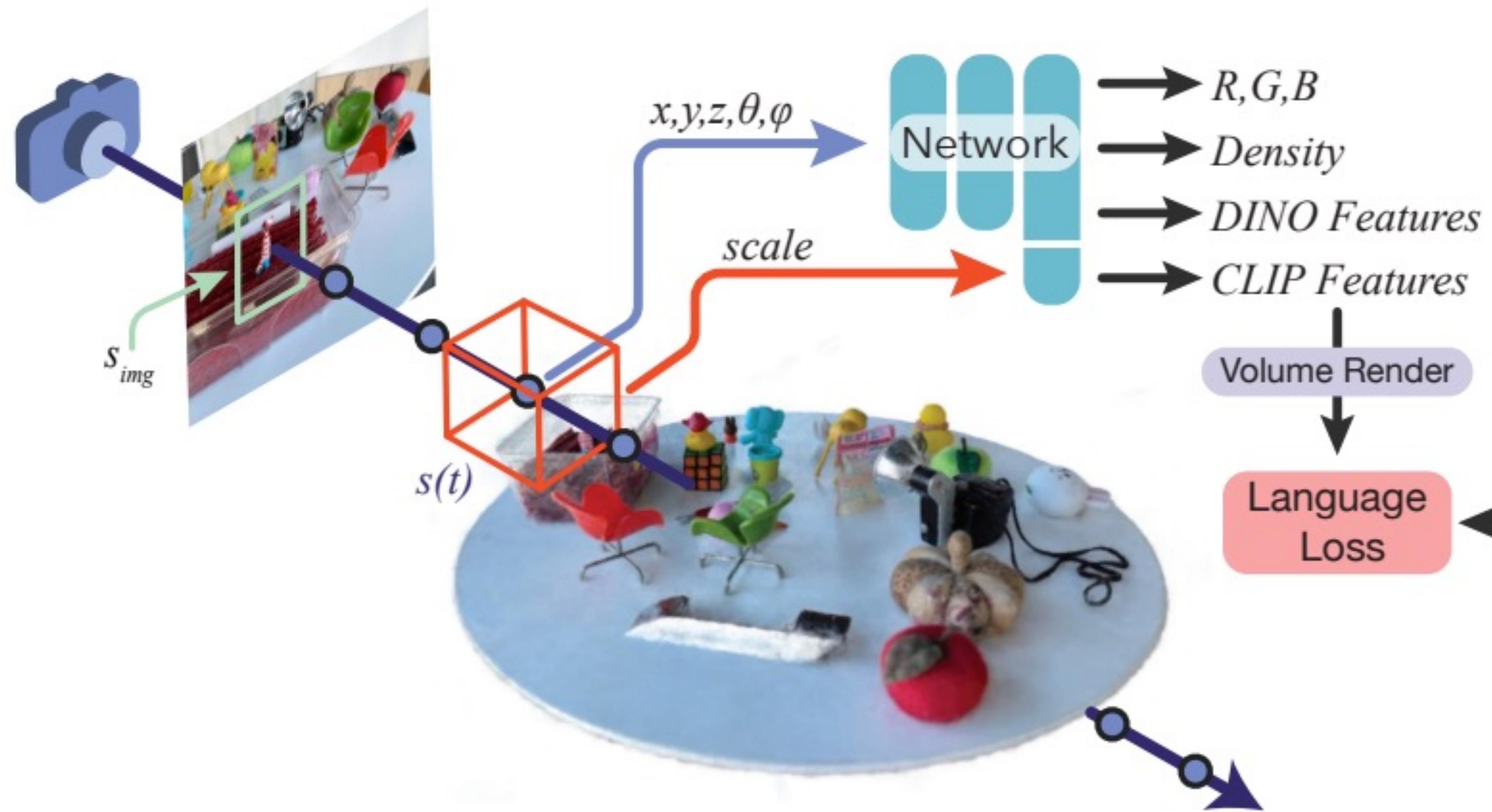


Figure 1: **Language Embedded Radiance Fields (LERF).** LERF grounds CLIP representations in a dense, multi-scale 3D field. A LERF can be reconstructed from a hand-held phone capture within 45 minutes, then can render dense relevancy maps given textual queries interactively in real-time. LERF enables a broad range of concepts to be queried via natural language, from abstract queries like “Electricity”, visual properties like “Yellow”, long-tail objects such as “Waldo”, and even reading text like “Boops” on the mug. For each prompt, an RGB image and relevancy map are rendered focusing on the location with maximum relevancy activation.

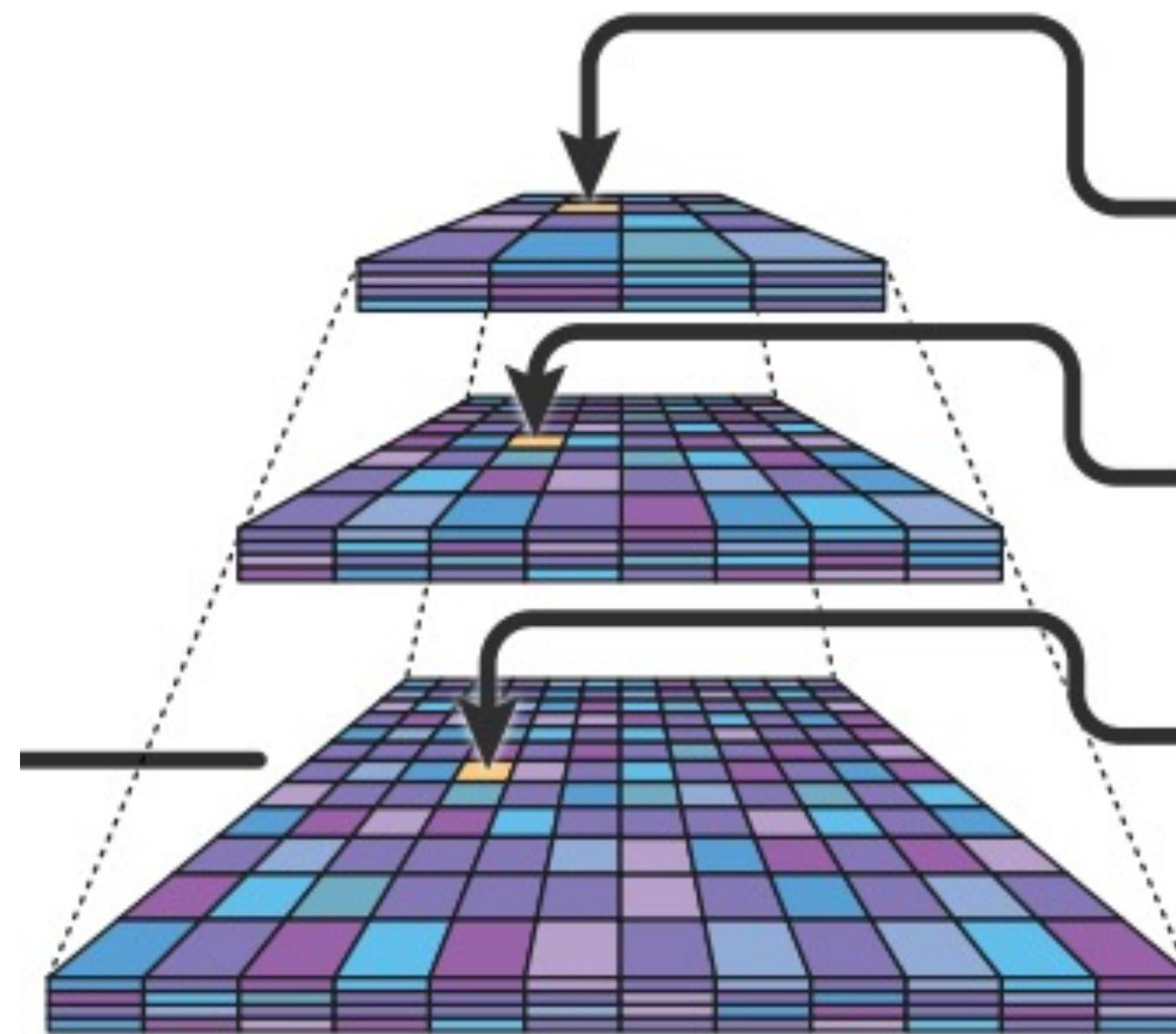


## LERF Rendering





## Multiscale CLIP Preprocessing



Multiscale CLIP Features

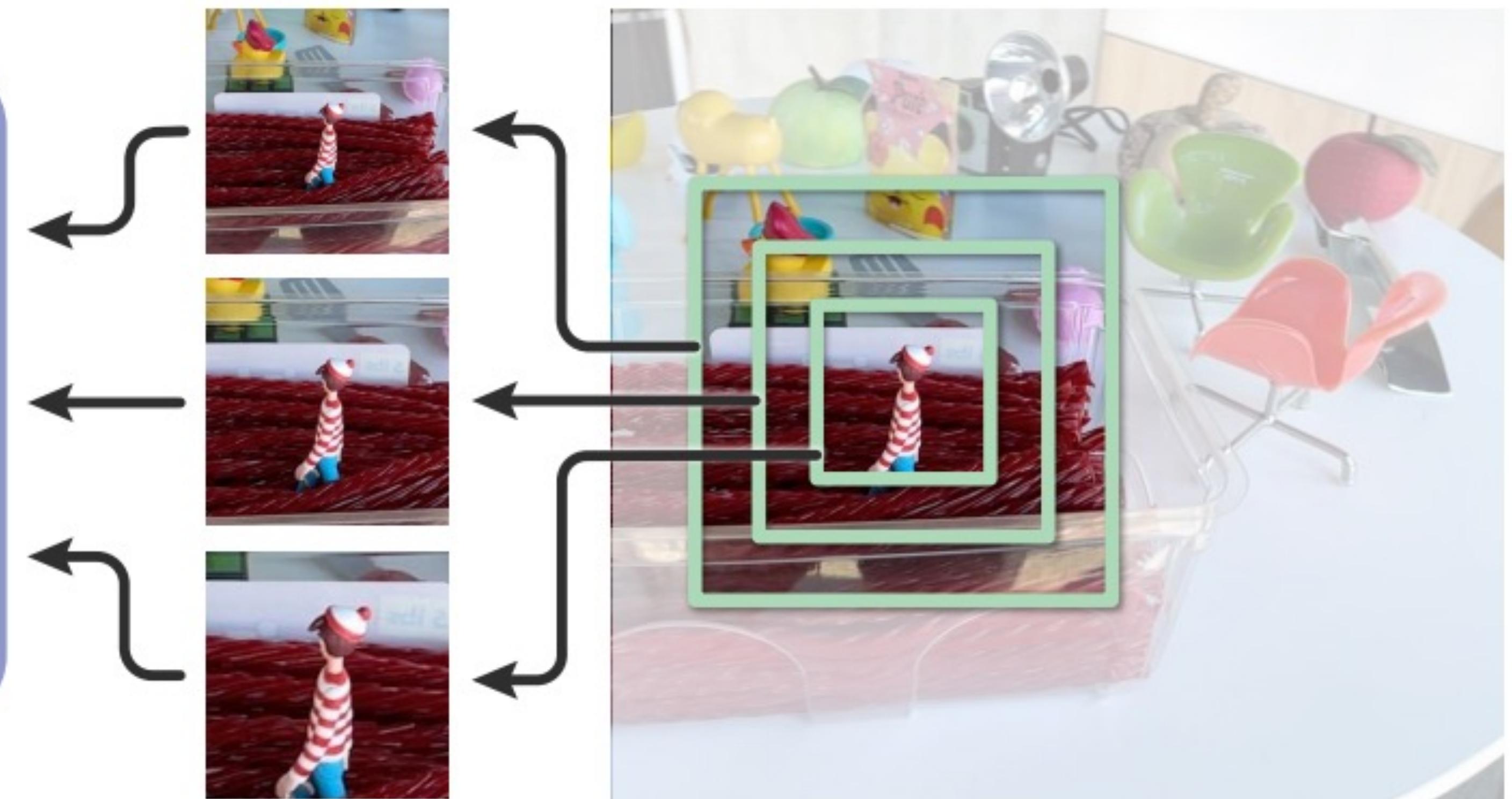
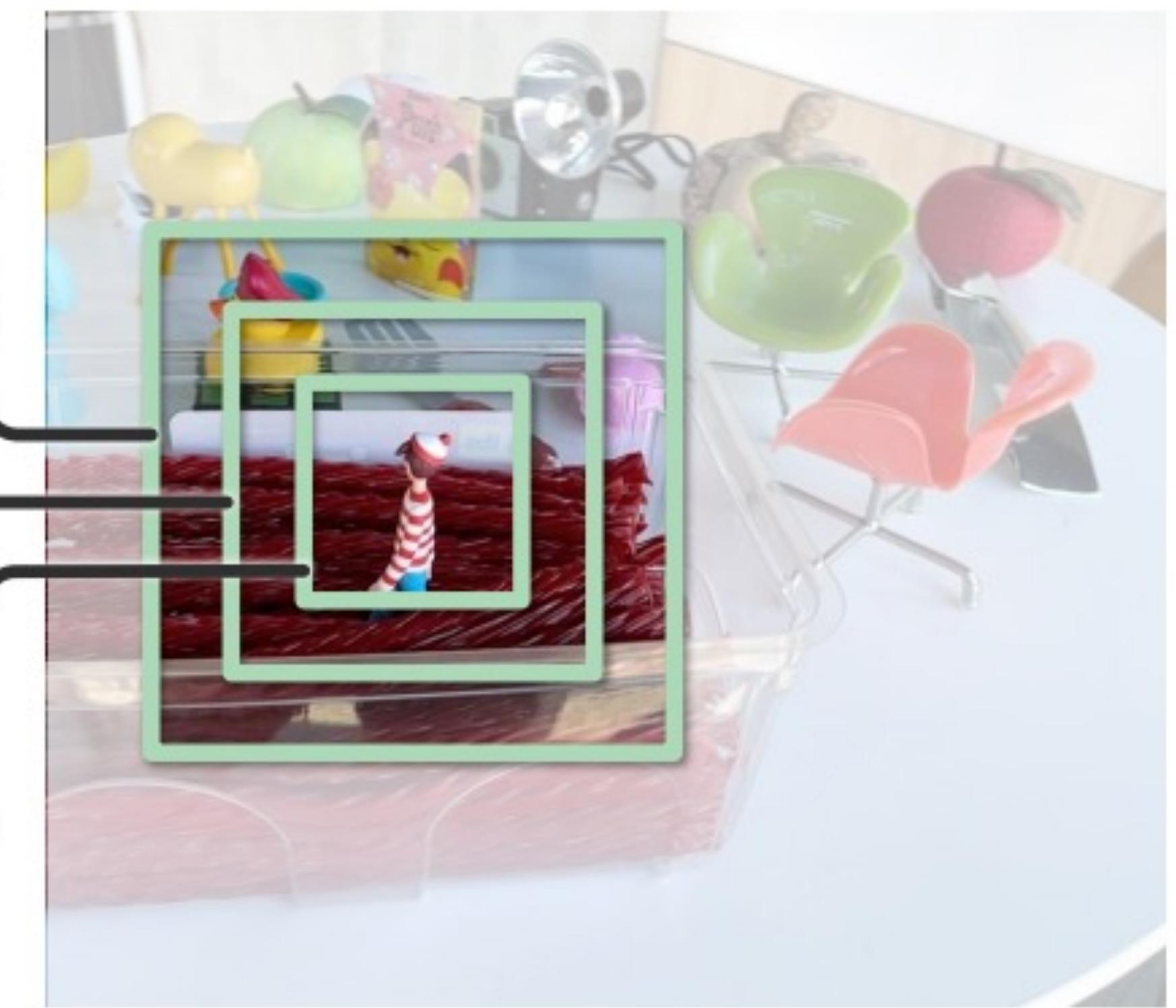


Image Patches



Training Image



# Putting it together

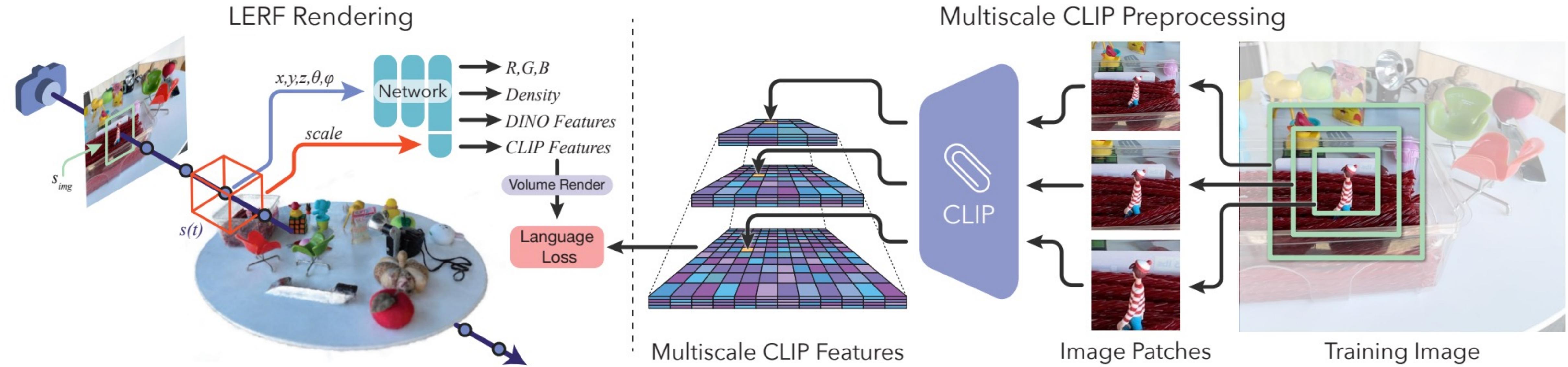
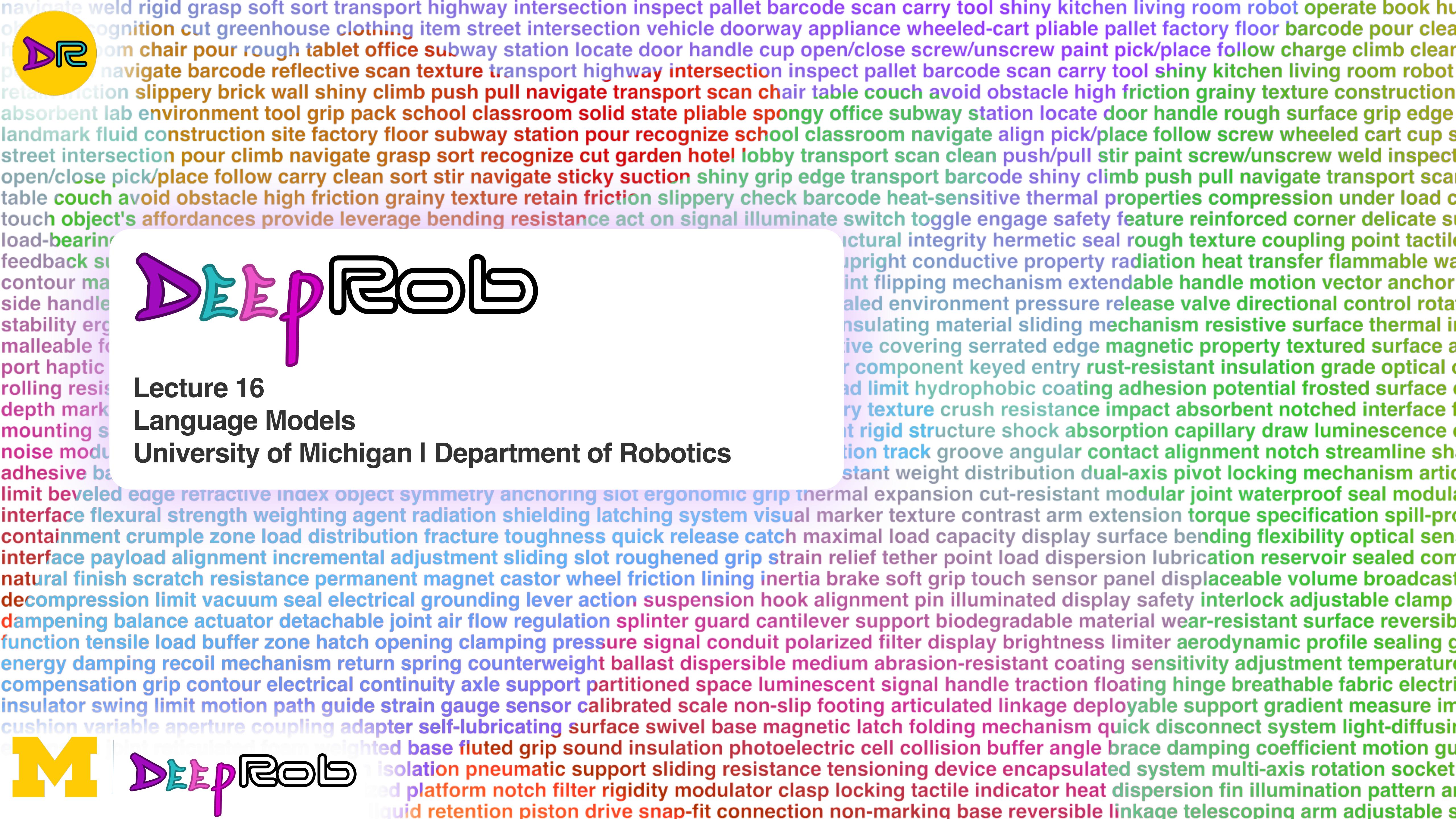


Figure 2: **LERF Optimization:** *Left:* LERF represents a field of 3D volumes, parameterized by position  $x, y, z$  and scale  $s$  (orange cube). To render a CLIP embedding along a ray, the field is sampled and averaged according to NeRF's volume rendering weights. Physical scale corresponds to an image scale  $s_{img}$  via projective geometry. *Right:* We pre-compute a multi-scale feature pyramid of CLIP embeddings over training views, and during training interpolate this pyramid with  $s_{img}$  and the ray's pixel location to obtain CLIP supervision. The CLIP loss maximizes cosine similarity, and other outputs are supervised with mean squared-error using standard per-pixel rendering.



weld rigid grasp soft sort transport highway intersection inspect pallet barcode scan carry tool shiny kitchen living room robot operate book h  
ognition cut greenhouse clothing item street intersection vehicle doorway appliance wheeled-cart pliable pallet factory floor barcode pour clea  
chair pour rough tablet office subway station locate door handle cup open/close screw/unscrew paint pick/place follow charge climb clear  
navigate barcode reflective scan texture transport highway intersection inspect pallet barcode scan carry tool shiny kitchen living room robot  
slippery brick wall shiny climb push pull navigate transport scan chair table couch avoid obstacle high friction grainy texture construction  
absorbent lab environment tool grip pack school classroom solid state pliable spongy office subway station locate door handle rough surface grip edge  
landmark fluid construction site factory floor subway station pour recognize school classroom navigate align pick/place follow screw wheeled cart cup s  
street intersection pour climb navigate grasp sort recognize cut garden hotel lobby transport scan clean push/pull stir paint screw/unscrew weld inspect  
open/close pick/place follow carry clean sort stir navigate sticky suction shiny grip edge transport barcode shiny climb push pull navigate transport scar  
table couch avoid obstacle high friction grainy texture retain friction slippery check barcode heat-sensitive thermal properties compression under load c  
touch object's affordances provide leverage bending resistance act on signal illuminate switch toggle engage safety feature reinforced corner delicate s  
load-bearing  
feedback su  
contour ma  
side handle  
stability erg  
malleable fo  
port haptic  
rolling resis  
depth mark  
mounting s  
noise modu  
adhesive ba  
limit beveled  
edge retractive  
inax object symme  
ancnoring slot ergonomic grip tne  
interface flexural strength weighting agent radiation shielding latching system visual marker texture contrast arm extension torque specification spill-pr  
containment crumple zone load distribution fracture toughness quick release catch maximal load capacity display surface bending flexibility optical sen  
interface payload alignment incremental adjustment sliding slot roughened grip strain relief tether point load dispersion lubrication reservoir sealed com  
natural finish scratch resistance permanent magnet castor wheel friction lining inertia brake soft grip touch sensor panel displaceable volume broadcast  
decompression limit vacuum seal electrical grounding lever action suspension hook alignment pin illuminated display safety interlock adjustable clamp  
dampening balance actuator detachable joint air flow regulation splinter guard cantilever support biodegradable material wear-resistant surface reversib  
function tensile load buffer zone hatch opening clamping pressure signal conduit polarized filter display brightness limiter aerodynamic profile sealing g  
energy damping recoil mechanism return spring counterweight ballast dispersible medium abrasion-resistant coating sensitivity adjustment temperatur  
compensation grip contour electrical continuity axle support partitioned space luminescent signal handle traction floating hinge breathable fabric electric  
insulator swing limit motion path guide strain gauge sensor calibrated scale non-slip footing articulated linkage deployable support gradient measure im  
cushion variable aperture coupling adapter self-lubricating surface swivel base magnetic latch folding mechanism quick disconnect system light-diffusin  
ighted base fluted grip sound insulation photoelectric cell collision buffer angle brace damping coefficient motion gu  
isolation pneumatic support sliding resistance tensioning device encapsulated system multi-axis rotation socket  
ed platform notch filter rigidity modulator clasp locking tactile indicator heat dispersion fin illumination pattern an  
quid retention piston drive snap-fit connection non-marking base reversible linkage telescoping arm adjustable s



# DEEP Rob

## Lecture 16

### Language Models

#### University of Michigan I Department of Robotics



DEEP Rob